

# Exotic Pest Information Collection and Analysis (EPICA) – gathering information on exotic pests from the World Wide Web\*

M. Bateman<sup>1</sup>, C. Brammer<sup>1</sup>, C. Thayer<sup>1</sup>, H. Meissner<sup>1</sup> and W. Bailey<sup>2</sup>

<sup>1</sup>Center for Integrated Pest Management, 1730 Varsity Drive, Suite 110, Raleigh, NC, 27606 (USA); e-mail: [Melanie.L.Bateman@aphis.usda.gov](mailto:Melanie.L.Bateman@aphis.usda.gov), [Colin.A.Brammer@aphis.usda.gov](mailto:Colin.A.Brammer@aphis.usda.gov), [Charles.L.Thayer@aphis.usda.gov](mailto:Charles.L.Thayer@aphis.usda.gov) and [Heike.E.Meissner@aphis.usda.gov](mailto:Heike.E.Meissner@aphis.usda.gov).

<sup>2</sup>USDA-APHIS-PPQ-Center for Plant Health Science and Technology, Plant Epidemiology and Risk Analysis Laboratory, 1730 Varsity Drive, Suite 300, Raleigh, NC 27606-5202 (USA); e-mail: [Woodward.D.Bailey@aphis.usda.gov](mailto:Woodward.D.Bailey@aphis.usda.gov)

This paper describes the Exotic Pest Information Collection and Analysis (EPICA) project, a cooperative effort of the United States Department of Agriculture's (USDA) Center for Plant Health Science and Technology (CPHST) and the National Science Foundation's Center for Integrated Pest Management (CIPM). The EPICA team identifies, archives, evaluates, and communicates relevant information about exotic plant pests from around the world that potentially threaten US agriculture and the environment. The goal of the EPICA project is to provide key groups within USDA's Plant Protection and Quarantine (PPQ) with regular updates of vital pest information to enable a proactive safeguarding approach.

## Introduction

International commerce and travel continue to increase worldwide. The United States' total agricultural imports grew from 23 billion USD in 1990 to 59 billion USD in 2005 (USDA, 2006a). The number of international visitors arriving in the US increased from 45 million in 1994 to 49 million in 2005 (Office of Travel, 2006a), and the number of US citizens travelling abroad grew from 27 million in 1996 to over 38 million in 2005 (Office of Travel, 2006b). This rise in global commerce and travel is accompanied by an increase in pest interceptions at US ports of entry (from 47 504 in 1990 to 64 679 in 2005) (USDA 2006b). If current trends in trade and species invasions continue unabated, it is projected that international commerce will lead to the establishment of several hundred exotic plant pest species in the US over the next 20 years (Levine & d'Antonio, 2003).

The US government, farmers, and consumers lose billions of dollars every year due to the negative impacts of introduced exotic species (Wheelis *et al.*, 2002). Once pest species become established over large areas, eradication is often impossible to achieve (Rejmánek & Pitcairn, 2002). Thus, a proactive approach including early warning systems that facilitate the exclusion of potential invaders is the preferred safeguarding strategy. A proactive approach may consist of many different elements including phytosanitary inspections focusing on high-risk commodities; ranking of the threat level of exotic pests; targeted pest surveys for detecting incipient populations;

or the development of mitigation and control strategies against imminent pest threats. Regardless of the specific components of the proactive approach, it critically depends on current and reliable pest information.

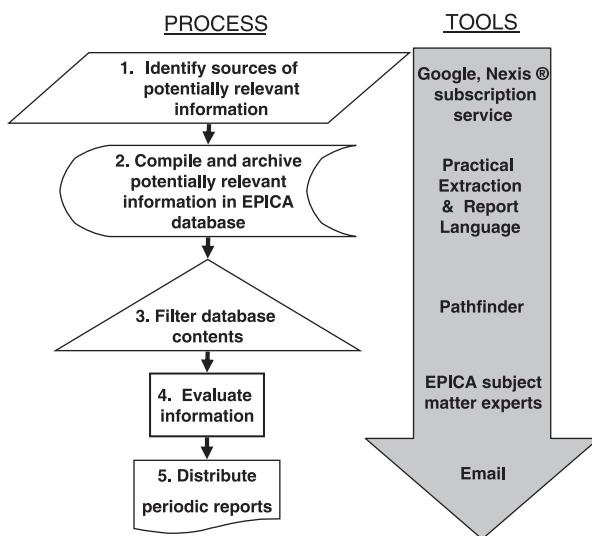
Information published in books and scientific journals, while generally reliable, is often published too late for safeguarding purposes. Meanwhile, the World Wide Web (WWW) is a valuable resource that includes abundant up-to-date information as well as much of the information available in printed media. The WWW consists of hundreds of petabytes (1000 000 000 000 000 bytes) of information (Gilder, 2006), is comprised of over 11 billion web pages (Gulli & Signorini, 2005), and continues to expand at a remarkable rate (Wilson, 2004).

Taking full advantage of the information available on the WWW presents an immense challenge on many different fronts. For example, the information is published in many different languages. Only a tiny fraction of the published information is relevant for safeguarding purposes. Of that, only a portion is accurate and sufficiently explicit. Some information is only available on a subscription basis or is restricted to certain user groups.

While Plant Protection and Quarantine (PPQ) employees have historically mined the WWW for information, this information is generally extracted in a piecemeal and uncoordinated fashion. Most users are limited by the search tools accessible to them (usually basic search engines), limited knowledge of foreign languages, and time. When valuable information is uncovered, it does not necessarily get distributed to all interested parties nor is it always archived in a central location.

Exotic Pest Information Collection and Analysis (EPICA) is a new PPQ project with a two-fold purpose:

\*Paper presented at the EPPO Conference on 'Computer Aids for Plant Protection' in Wageningen, the Netherlands, 2006-10-17/19.



**Fig. 1** The five-step process used by EPICA to collect and distribute information on exotic plant pests that potentially pose a threat to US agriculture.

- 1 to create a streamlined, efficient process for continuously collecting, archiving, and analyzing relevant information about exotic plant pests and
- 2 to effectively distribute this information to the appropriate recipients within PPQ.

The EPICA project has a dedicated team of subject matter experts who have at their disposal powerful data mining tools, as well as foreign language expertise that covers almost 90% of all Internet content.

The EPICA team is presently testing and using for its purposes the following five-step process (Fig. 1) developed by the USDA Veterinary Services' Center for Emerging Issues:

- 1 Sources of potentially relevant information are identified
- 2 Information provided by these sources is continuously compiled and archived in a database
- 3 Boolean logic and key words are used to selectively extract information from this database
- 4 The relevance and accuracy of this information is carefully evaluated by the EPICA team
- 5 Summary reports of the relevant information are periodically produced. The following section explains this process in more detail.

## The process

### 1. Identifying information sources

The EPICA team identifies potentially relevant electronic information sources (e.g. webpages, listservs, weblogs, e-journals, news services). This is accomplished by using conventional search engines (e.g. Google) as well as Lexis® (LexisNexis Group, 2006). Lexis® is a paid service providing access to some information that is only available on a subscription basis. It offers over 32 000 information sources, including newspapers, trade journals, newsletters, etc.

Each of the identified sources of potentially relevant information is recorded along with details about the resource type, the relative value of its content, its geographic scope, the frequency with which it is updated, the language in which it is published, and the identity of the publisher.

To date, the EPICA team has identified over 1000 regularly updated electronic sources of potentially useful information, authored by agribusiness and trade organizations, academic researchers and professional societies, nongovernmental organizations and informal clubs, government agencies, and news organizations. These sources originate from 126 countries and are published in 57 different languages.

### 2. Compiling and archiving information

Information is automatically retrieved on a daily basis from each of the potentially relevant sources in a nonselective fashion using Practical Extraction and Report Language (PERL). This information is stored in a database in the form of text documents.

Between May and October of 2006, a total of 16 405 text documents were collected and stored in the database. The rate at which documents are added to the database has increased exponentially, and it is anticipated to continue growing dramatically as the EPICA project progresses.

### 3. Filtering database contents

In order to continuously extract the most pertinent information from the database, the EPICA team uses Pathfinder, a powerful data mining and visualization tool developed by the US military (Pathfinder Version 5.2.2, 2004). This software allows the use of queries (key words in combination with Boolean logic) to filter the database contents. It also provides tools for query building (e.g. customizable thesaurus for key words), browsing of query results (e.g. highlighting of keywords and query terms), and visualization of database contents (e.g. histograms, timelines, a graphical filing system).

### 4. Evaluating information

After using Pathfinder to filter the information in the project database, the EPICA team reads through the extracted articles, evaluating their reliability, timeliness and significance.

### 5. Reporting information

In coordination with prospective (PPQ-internal) clients, the EPICA team is currently developing a system by which summary reports will be distributed periodically by e-mail.

## Closing remarks

By building a solid infrastructure, the EPICA team is making strides toward its goal of providing PPQ with up-to-date information on exotic plant pests.

As invasive species are a global problem, communication and collaboration between nations can help to address this issue. The EPICA team seeks partners who share the same goals. Because much of the relevant information on pests that potentially threaten agriculture is freely available on the WWW, the EPICA strategy can be implemented by other countries.

## Acknowledgements

We thank Elizabeth Williams of USDA-CEI for providing technical support for the EPICA database; Tony Koop of USDA-CPHST for providing useful data; Andrea Lemay for proofreading an earlier version of this manuscript; the CIPM team of the Global Pest and Disease Database team for providing pest lists and query terms. Funding for this project was provided by the USDA-APHIS Office of Emergency Management and Homeland Security.

## **Analyse et collecte d'informations sur les organismes nuisibles exotiques (EPICA) – Rassembler des informations à partir d'Internet**

Cet article décrit le projet EPICA (Exotic Pest Information Collection and Analysis), un projet coopératif émanant du Centre pour la science et les technologies phytosanitaires (CPHST) de l'USDA (United States Department of Agriculture) et la Fondation nationale pour la recherche sur la protection intégrée (National Science Foundation's Center for Integrated Pest Management – CIPM). L'équipe EPICA identifie, archive, évalue et transfère les informations pertinentes sur les organismes exotiques nuisibles aux végétaux du monde entier qui menacent potentiellement l'agriculture et l'environnement des Etats-Unis. Le but du projet EPICA est de fournir aux personnes clés au sein du département Protection des Plantes et Quarantaine (PPQ) de l'USDA des mises à jour régulières sur les informations cruciales relatives aux organismes pour permettre une approche de sauvegarde proactive.

## **Сбор и анализ информации об экзотических вредных организмах (EPICA). Сбор информации об экзотических вредных организмах во Всемирной Паутине**

В статье дается описание проекта Сбора и анализа информации об экзотических вредных организмах (EPICA) –

совместной работы Департамента сельского хозяйства США (USDA), Центра по науке и технологии фитосанитарии (CPHST) и Центра Национального научного фонда по Интегрированному управлению вредными организмами (CIPM). Команда EPICA находит, архивирует, дает оценку и передает соответствующую информацию со всего мира об экзотических вредных организмах, которые могут угрожать сельскому хозяйству и экологии США. Целью проекта EPICA является предоставление ключевым группам в отделе Зоотехники и карантине растений (PPQ) в рамках USDA регулярно обновляемой наиболее важной информации о вредных организмах, с тем чтобы обеспечивать упреждающий подход к сохранению сельского хозяйства и экологии.

## References

- Gilder G (2006) The information factories. *Wired Magazine* **14**, 1–5.
- Gulli A & Signorini A (2005) The indexable web is more than 11.5 billion pages. *Proceedings of the WWW 2005 Conference*. <http://www2005.org/cdrom/docs/p902.pdf> [accessed on 17 October 2006].
- Levine JM & D'Antonio CM (2003) Forecasting biological invasions with increasing international trade. *Conservation Biology* **17**, 322–326.
- LexisNexis Group (2006) Nexis®. Dayton, Ohio (US). <http://nexis.com> [accessed on 17 October 2006].
- Office of Travel & Tourism Industries (2006a) *US International Air Travel Statistics (I-92 data)*. <http://tinet.ita.doc.gov/view/m-2005-O-001/index.html> [accessed on 17 October 2006].
- Office of Travel & Tourism Industries (2006b) *Visitor Arrivals Program (I-94 Form)*. <http://tinet.ita.doc.gov/view/m-2005-I-001/index.html> [accessed on 17 October 2006].
- Pathfinder (2004) *Pathfinder Version 5.2.2*. Exploitation Technologies Division, SAIC, Charlottesville, Virginia (US).
- Rejmánek M & Pitcairn MJ (2002) When is eradication of exotic pest plants a realistic goal? *Turning the Tide: the Eradication of Invasive Species* (eds Veitch, CR & Clout, MN), pp. 249–253. IUCN SSC Invasive Species Specialist Group, IUCN, Gland (CH) and Cambridge (GB).
- United States Department of Agriculture (USDA) (2006a) *Foreign Agricultural Trade of the United States, Economic*. Research Service, <http://www.ers.usda.gov/Data/FATUS/DATA/Cynonag.xls> [accessed on 17 October 2006].
- United States Department of Agriculture (USDA) (2006b) *Agricultural Quarantine Activity System: PestID, Version 1.2*. Washington D.C. (US) [accessed on 18 October 2006].
- Wheelis M, Casagrande R & Madden L (2002) Biological attack on agriculture: Low-tech, high impact bioterrorism. *Bioscience* **52**, 569–576.
- Wilson EJ III (2004) *The Information Revolution and Developing Countries*. MIT Press, Cambridge, Massachusetts (US).